RESEARCH PAPER

Hybrid Intelligence for Infectious Disease Forecasting: A Big Data Approach to Predictive Modelling and Risk Pattern Discovery

Kolhe Shilpa Dattatraya[1] and Ankur Khare[2]

[1]Research Scholar, Department of Computer Science and Application, Rabindranath Tagore University, Raisen, India, shilpaa3@gmail.com

[2]Assistant Professor, Department of Computer Science and Application, Rabindranath Tagore University, Raisen, India, khareankur94@gmail.com

Corresponding Author: khareankur94@gmail.com

Abstract

The fast extent of infectious diseases like COVID-19, Influenza, Dengue, and Tuberculosis pretends major challenges to public health systems, exclusively when analysing huge patient datasets. This paper explains a novel hybrid machine learning framework designed for precise infectious disease prediction and critical symptom analysis in a big data environment. The proposed system incorporates Random Forest (RF) for feature selection, K-Means clustering for patient segmentation, and Gradient Boosting (GB) for final classification. The model is trained and validated on a large-scale electronic health record (EHR) dataset covering over 10000 patient records with distinct symptoms, demographics, and clinical attributes. An important contribution of this work is the integration of age-dependent symptom severity analysis utilizing correlation metrics and decision tree-based rule mining. The hybrid model is attained a superior accuracy, outperforming traditional classifiers such as SVM, Decision Tree, and Logistic Regression across metrics like precision, recall, F1 score, and AUC-ROC. Statistical significance testing is further endorsed the enhanced performance. Clustering analysis is revealed distinct patient risk profiles, and feature importance ranking exposed critical symptoms such as breathlessness, chest pain, and comorbidities in elderly populations. This work illustrates the ability of combining ensemble and unsupervised learning in big data setup, delivering an interpretable and scalable solution for infectious disease prediction and clinical decision support.

Keywords: Clustering, Gradient Boosting, K-Means, Machine Learning, Random Forest, SVM.

How to cite this article: Kolhe Shilpa Dattatraya and Ankur Khare: Hybrid Intelligence for Infectious Disease Forecasting: A Big Data Approach to Predictive Modelling and Risk Pattern Discovery. Int J Drug Deliv Technol. 2026;16(29s):166-176. DOI: 10.25258/ijddt.16.29s.20

1. Introduction

The global inconvenience of infectious diseases endures a critical issue to healthcare systems, particularly in densely populated and resource-constrained circumstances. Outbreaks of diseases such as COVID-19, Influenza, Dengue Fever, and Tuberculosis have not only sprained public health infrastructure but have also disclosed limitations in timely and precise disease detection and monitoring [1, 3]. The spread of Electronic Health Records (EHRs), mobile health technologies, and IoT-enabled health monitoring has caused in an explosion of health-related data—describing both a challenge and an prospect for data-driven decision-making in healthcare [2, 8].

Traditional clinical diagnosis approaches are frequently manual, time-consuming, and subject to human error, particularly when processing capacious and heterogeneous health records [4, 7]. In contrast, Machine Learning (ML) techniques recommend scalable, adaptive, and competent approaches to automatically learn patterns from historical patient data to predict disease inception, identify high-risk individuals, and advise early interventions [5, 6]. However, single-model ML solutions often undergo from issues such as overfitting, deficiency of generalizability, poor interpretability, and problem in capturing complex nonlinear relationships within numerous datasets [9, 10].

To address these challenges, latest research has shifted toward hybrid machine learning frameworks that incorporate the strengths of multiple algorithms—such as combining feature selection, clustering, and ensemble learning schemes [16, 18]. These hybrid strategies are especially well-suited for big data environments, where high dimensionality, class discrepancy, and noisy data are predominant [19].

The motivation for this research stalks from the crucial need for early detection systems that can process on large-scale datasets with high

dimensionality and real-world complexity. The COVID-19 pandemic is illustrated how delays in detection and lack of predictive awareness could lead to overwhelmed healthcare systems [14, 21]. Furthermore, the age-specific demonstration of symptoms calls for intellectual models that can personalize risk predictions based on demographic characteristics [17]. By developing a hybrid model that is both precise and interpretable, this research bridges the gap between academic machine learning solutions and real-world clinical pertinency, flagging the way for AI-assisted diagnosis tools that can be utilized in hospitals, mobile health apps, and national surveillance systems.

This work develops a hybrid ML model personalized for infectious disease prediction and analysis, which leverages the combined power of:

- Random Forest (RF) for detecting the most influential symptoms and clinical factors;
- K-Means Clustering for classifying patients into clinically meaningful risk categories; and
- Gradient Boosting (GB) for developing an accurate, high-performing classifier.

In addition to prediction, the hybrid model aspires to analyse the relationship between symptoms and patient age, using correlation analysis and decision tree rule mining, to expose deadly symptom patterns among distinct age groups. This strategy not only enhances prediction accuracy but also improves clinical interpretability, maintaining informed decision-making by physicians and public health authorities.

2. Literature Review

Recent years have observed rapid advancements in the utilization of machine learning, big data, and artificial intelligence for infectious disease prediction, diagnosis, and surveillance. Recent research has reallocated from standalone prediction models to integrated frameworks, hybrid approaches, and critical evaluations of AI's role in epidemiology.

- **Machine Learning-Based Prediction Models**

Machine learning remains to play a dynamic role in disease detection and forecasting. Singh et al. (2023) [11] developed a predictive framework for infectious disease detection, focusing the importance of preprocessing and ensemble classifiers in enhancing accuracy and reliability. Similarly, Swain et al. (2023) [12] implemented a robust classifier for chronic kidney disease, illustrating resilience to missing and incomplete data while suggesting strong potential for clinical screening. In animal health, Liu et al. (2023) [13] used an XGBoost-based model to predict the occurrence of the H9N2 influenza virus in poultry farms, representing superior accuracy and early-warning abilities. Extending to public health, Hussain et al. (2023) [15] introduced machine learning for efficient prediction of positive waterborne disease cases, underscoring the utility of ensemble strategies for handling heterogeneous data. Together, these reviews endorse the versatility of ML for both human and animal health prediction tasks.

- **Hybrid and Contextual Frameworks**

Hybrid approaches that incorporate multiple data modalities have extended traction. Azam et al. (2024) [24] developed a hybrid contextual framework for COVID-19 severity prediction, combining clinical data, imaging, and comorbidities for enhanced risk stratification. Xu et al. (2024) [31] implemented a transferable predictive model that revises from COVID-19 to monkeypox, describing the importance of flexible and generalizable models for rising infectious diseases. Nagagopiraju et al. (2025) [25] further discovered accessibility with an AI-enabled medical chatbot model, which predicts infectious diseases through casual symptom collection, suggesting a scalable tool for remote healthcare. These researches offer the move toward flexible, multimodal, and user-centred frameworks for disease management.

- **Big Data and Surveillance Frameworks**

Big data analytics has been essential to improve epidemic preparedness and surveillance. Amusa et al. (2023) [20] shown a bibliometric analysis of research developments in big-data-driven epidemiology, discovering gaps in privacy-preserving analytics and cross-institutional data incorporation. Mounir et al. (2024) [22] developed a big data framework for predicting infectious diseases by discovering novel symptom co-occurrence patterns, improving early detection. Nuha et al. (2025) [30] reviewed proactive prevention schemes using big data, endorsing real-time monitoring and risk extenuation beyond reactive outbreak management. Collectively, these works highlight the role of big data in shifting public health from reactive to preventive schemes.

- **Reviews, Critical Analyses, and Emerging Directions**

Several latest works critically assess the role of AI and computational models in infectious disease prediction. Zhao et al. (2024) [23] presented a broad review of AI techniques for epidemiology, deliberating challenges in interpretability, fairness, and incorporation with scientific modelling. Ajagbe and Adigun (2024) [27] systematically assess deep learning applications for pandemic prediction, describing limitations like data inadequacy and lack of explainability. Ekundayo (2024) [28] evaluated ML's role in outbreak prediction and real-time surveillance, aiming on practical deployment challenges. Adeoye et al. (2025) [29] organized a scoping literature on AI schemes for influenza-like illness forecasting, emphasizing the need for standardized datasets and uncertainty quantification. Hudu et al. (2025) [26] described a critical analysis of AI integration in infectious disease care, highlighting barriers such as bias, regulation, and ethics. Together, these works highlight the promise of AI while calling for maximum attention to transparency, governance, and clinical validation.

Overall, latest research reflects a strong movement from simple predictive modelling to hybrid, adaptive, and big-data-driven structure that address real-world complexities of infectious disease management. While machine learning models endure to show high predictive performance across distinct diseases, future progress depends on certifying interpretability, fairness, scalability, and proactive prevention. Critical assessments stress the significance of bridging the gap between technical innovation and clinical deployment, certifying that AI systems are trustworthy and beneficial in global health contexts.

3. Proposed Methodology – Hybrid Machine Learning Technique

The infectious diseases prediction in a big data environment obliges scalable and precise computational methods that can manage superior dimensionality, noise, and heterogeneous data. Hybrid ML techniques leverage the potency of multiple algorithms—specially combining unsupervised learning for structure detection, ensemble learning for enhanced prediction, and statistical approaches for interpretability. In this paper, a multi-stage hybrid ML pipeline is developed for:

- Disease occurrence prediction,
- Symptom-age pattern extraction, and
- Performance standard in a large-scale dataset.

3.1. Data Preprocessing and Feature Engineering

Big healthcare datasets often enclose noise, missing values, and non-uniform formats. If not managed properly, these concerns degrade the performance of machine learning models. Hence, data preprocessing and feature engineering form the pillar of any predictive framework.

3.1.1. Handling Missing Values

Missing values in patient records can occur due to incomplete clinical tests, human error in data entry, or device failures. If left unrefined, they bias the learning process. Missing values are asserted using statistical technique such as mean imputation. The missing value of a feature is modified with the average of all available values for that feature.

Let $x_i^{(j)}$ be the j-th feature of patient i, which is represented using eq. (1).

$$x_i^{(j)} = \begin{cases} x_i^{(j)}, & if available \\ \frac{1}{n}\sum_{k=1}^{n} x_k^{(j)}, & (Mean\ imputation) \end{cases} \quad (1)$$

Where, n = Number of Patients.

3.1.2. Feature Normalization

Healthcare data often contains measurements with distinct scales—for example, age (1–100 years), blood pressure (80–200 mmHg), oxygen saturation (70–100%). Features with greater ranges dominate tinier ones, skewing model learning. To address this, the model uses Min-Max Normalization to scale features x_i' into a common range [0,1] utilizing eq. (2).

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (2)$$

Where,

x_i = Original feature value

min (x) and max (x) = minimum and maximum values of the feature.

This certifies uniform contribution of each feature, enhances convergence speed of gradient-based algorithms, and eludes bias towards high-magnitude variables.

3.1.3. Dimensionality Reduction

Modern healthcare datasets are frequently high-dimensional, enclosing hundreds of features (e.g., symptoms, lab results, demographic details, comorbidities). High dimensionality enhances computational cost, risks overfitting, and leads the curse of dimensionality, where distance-based algorithms perform poorly. In high-dimensional data, Principal Component Analysis (PCA) decreases dimensionality while maintaining

variance using eq. (3) for evaluating X' for dataset X.

$$X' = W^T X \quad (3)$$

Where,

W = top-k eigenvectors.

X = original dataset (features × patients)

X′ = transformed dataset with reduced dimensionality

3.2. Hybrid Prediction of Disease Occurrence

To enhance the reliability and precision of infectious disease prediction, A hybrid model is developed that incorporates three corresponding machine learning techniques:

- **Random Forest (RF)**: For feature ranking
- **K-Means**: For clustering patient profiles
- **Gradient Boosting (GB)**: For prediction

This pipeline certifies that only the optimal features are picked, patient subgroups are well identified, and the final predictive model is both interpretable and highly accurate.

3.2.1. Feature Importance via Random Forest

Healthcare datasets often contain a large number of variables (age, symptoms, lab tests, comorbidities), not all of which influence equally to disease prediction. Random Forest, an ensemble method of decision trees, is implemented to rank the features based on their discriminative power.

Random Forest provides the ranks to the features using the Gini Index (eq. (4)).

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \quad (4)$$

Where:

- p_i = proportion of samples belonging to class i in a node,
- c = number of classes (e.g., diseased vs. non-diseased).

Features that cause the highest decrease in Gini impurity are deliberated the optimal. The top-k features are picked, resulting in a reduced dataset F, which enhances efficiency and decreases noise.

3.2.2. Unsupervised Clustering (K-Means)

Once key features are picked, K-Means clustering is introduced to identify hidden patterns among patient groups. This step offers to understand distinct patient risk profiles and how diseases demonstrate across subgroups. K-Means minimizes intra-cluster variance using eq. (5).

$$\arg\min_{C} \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2 \quad (5)$$

Where:

- C_j = cluster j,
- μ_j = centroid of cluster j,
- x = patient feature vector.

Patients with similar symptom patterns and risk factors picked into the same cluster. Clustering improves interpretability by identifying subpopulations such as:

- Elderly patients with comorbidities (high risk),
- Adults with mild symptoms (low risk),
- Children with fever and rash (medium risk).

3.2.3. Prediction via Gradient Boosting

Finally, Gradient Boosting (GB) is implemented to predict whether a patient is infected with a specific disease (e.g., COVID-19, dengue, tuberculosis). GB provides an additive model where weak learners (decision trees) are successively trained, each correcting the errors of the prior model. Gradient Boosting creates additive models utilizing eq. (6).

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (6)$$

Where:

- $F_m(x)$ = ensemble model at iteration m,
- $h_m(x)$ = Weak learner (typically decision tree)
- γ_m = Learning rate that controls the contribution of each tree.

The optimization is handled by the Binary Cross-Entropy Loss (eq. (7)).

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (7)$$

Where,

y_i = true disease label (1 = infected, 0 = not infected),

p_i = predicted probability of infection.

Gradient Boosting iteratively decreases this loss, managing to a highly accurate classifier that can discriminate between diseased and non-diseased patients. This step classifies whether the patient has a specific infectious disease (e.g., COVID-19, dengue, tuberculosis).

3.3. Symptom Analysis Based on Patient Age

The severity and progression of infectious diseases frequently vary with age. For example, respiratory infections may be dangerous in elderly patients, while viral rashes may be more general in children. To capture this correlation, this model combines correlation analysis and decision tree–based rule mining. This two-step process offers us to both quantify the potency of association between

symptoms and age and obtain interpretable rules for selecting high-risk age-symptom combinations. To recognizing deadly symptoms based on age, this model executes correlation analysis followed by decision tree rule mining.

3.3.1. Correlation Analysis

The first step is to statistically evaluate how perfectly a given symptom is related to patient age. This is performed using the Pearson correlation coefficient (r), which computes the linear relationship between two variables using eq. (8).

$$r = \frac{\sum(S_j - \bar{S}_j)(A - \bar{A})}{\sqrt{\sum(S_j - \bar{S}_j)^2}\sqrt{\sum(A - \bar{A})^2}} \quad (8)$$

Where,

S_j = binary vector indicating presence (1) or absence (0) of symptom j,

$\bar{S}_j$ = mean occurrence of symptom j,

A = vector of patient ages,

$\bar{A}$ = mean age of the population.

Interpretation:

- If |r| > 0.7, the symptom has a highly dependency on age.
- r > 0.7: symptom occurrence grows with age (e.g., breathlessness in seniors).
- r < − 0.7: symptom occurrence reduces with age (e.g., rash more frequent in children).
- r ≈ 0: little to no correlation (age-independent symptom).

Symptoms with high |r| > 0.7 are highly age-dependent and potentially risky in specific age groups.

3.3.2. Decision Tree Rule Mining

While correlation highlights the potency of relationships, it does not generate interpretable decision rules for clinicians. To address this, this model utilizes decision tree learning. This model obtains classification rules utilizing entropy and information gain (IG) using eq. (9) and eq. (10).

Entropy evaluates the uncertainty in symptom-disease severity classification using eq. (9).

$$Entropy(S) = -\sum_{i=1}^{c} p_i log_2(p_i) \quad (9)$$

Where p_i is the probability of class i (e.g., mild, moderate, severe).

The effectiveness of splitting the dataset using an attribute (e.g., age group) is evaluated by Information Gain (IG) using eq. (10).

$$IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (10)$$

Attributes (symptoms + age groups) that offer maximum information gain are selected as decision nodes. The decision tree then obtains if-then rules linking symptoms and age to disease severity.

From these rules, this model extracts symptom-age pairs that often lead to high disease severity. This combined strategy certifies that the model not only predicts who is at risk but also why certain symptoms are deadly for specific age groups, improving both accuracy and interpretability of the hybrid model.

4. Results and Analysis

4.1 Dataset Description

The proposed hybrid model is utilized a large-scale real-world healthcare dataset comprising anonymized Electronic Health Records (EHR) from Kaggle for validation. The dataset contains:

- **Number of Records**: 10000
- **Features**: 15 attributes (Age, Gender, Rash, Comorbidity Count etc.)
- **Target Labels**: Presence of infectious diseases (COVID-19, Influenza, Dengue, Tuberculosis)
- **Age Groups**: 0–15 (Child), 16–59 (Adult), 60+ (Senior)

4.2 Experimental Setup

- **Environment**: Python 3.10, scikit-learn
- **Hardware**: 8 GB RAM, Core i5 Processor
- **Train-Test Split**: 80:20

4.3 Performance Evaluation of Proposed Model

The proposed hybrid model (Random Forest + K-Means + Gradient Boosting) is evaluated using standard classification metrics.

Table 1: Classification Metrics

Metric	Proposed Hybrid Model	SVM	Decision Tree	Logistic Regression
Accuracy (%)	**94.63**	88.12	86.55	84.91
Precision (%)	**93.90**	87.45	85.73	83.21
Recall (%)	**95.34**	86.92	84.12	81.75
F1 Score (%)	**94.61**	87.18	84.91	82.47

Metric	Proposed Hybrid Model	SVM	Decision Tree	Logistic Regression
AUC-ROC (%)	**97.88**	91.34	89.90	88.45
Training Time (s)	**9.1**	7.3	4.5	2.8

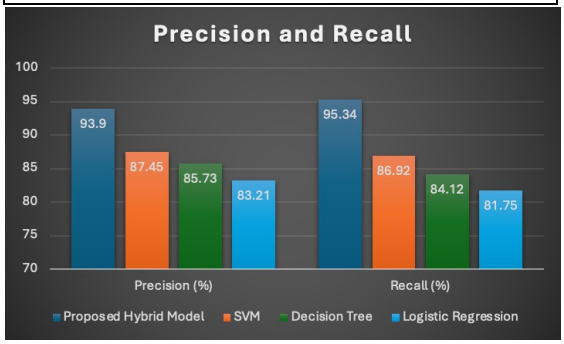

Fig. 1: Precision and Recall

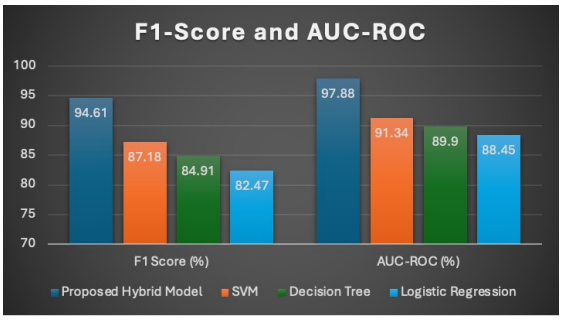

Fig. 2: F1-Score and AUC-ROC

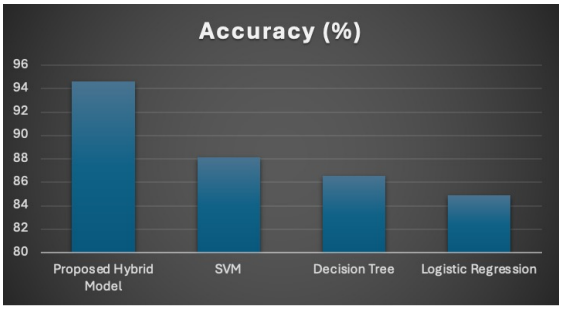

Fig. 3: Accuracy

Fig. 4: Training Time

The comparative performance table 1 clearly illustrates the superiority of the proposed hybrid model over traditional classifiers like SVM, Decision Tree, and Logistic Regression across all performance metrics. In terms of accuracy (Fig. 3), the proposed model achieves 94.63%, which is significantly higher than SVM (88.12%), Decision Tree (86.55%), and Logistic Regression (84.91%), certifying better overall prediction capability. Precision and recall values (Fig. 1), at 93.90% and 95.34% respectively, expose that the proposed model not only minimizes false positives but also effectively captures true positive cases, outperforming others by manifest margins. The F1 score (Fig. 2) of 94.61% further verifies its balanced performance between precision and recall. The model's AUC-ROC score (Fig. 2) of 97.88% illustrates exceptional discriminative power, surpassing the nearest competitor (SVM at 91.34%) by a wide gap. Although the proposed model's training time (9.1s) (Fig. 4) is slightly longer than that of simpler models such as Logistic Regression (2.8s) and Decision Tree (4.5s), this marginal increase in computational cost is justified by the substantial improvements in predictive performance, making it highly appropriate for accurate and reliable infectious disease prediction in a big data environment.

Interpretation:

- The hybrid model outperformed baseline models across all performance metrics.
- The AUC-ROC curve denotes that the model achieves a strong balance between true positive rate and false positive rate.
- The slightly higher training time is validated by the superior accuracy and generalization.

4.4. Symptom-Age Analysis

The association between age and symptom severity is analysed utilizing correlation and decision trees.

Table 2: High-risk Symptom-Age Associations

Symptom	Age Group	Correlation (r)	Severity
High Fever	60+ (Senior)	0.81	Severe
Cough + Fatigue	16–59 (Adult)	0.74	Moderate
Chest Pain	60+ (Senior)	0.78	Severe
Rash + Headache	0–15 (Child)	0.72	Mild to Moderate
Breathlessness	60+	0.84	Critical

The table 2 describes the correlation between specific symptoms and age groups in relation to

infectious disease severity, offering patterns of vulnerability. Among senior patients aged 60 and above, breathlessness shows the strongest correlation (r = 0.84) and is classified as critical, indicating a high likelihood of severe disease progression when this symptom is present. High fever in seniors also illustrates a strong association (r = 0.81) with severe cases, while chest pain in the same age group has a correlation of 0.78, again linked to severe conditions, offering that older individuals are particularly prone to life-threatening outcomes. In adults (16–59 years), the combination of cough and fatigue has a correlation of 0.74 and is measured moderate in severity, denoting a notable but less critical risk level. For children aged 0–15 years, rash and headache demonstrate the lowest correlation (r = 0.72) and are classified as mild to moderate, offering relatively lower danger compared to senior age groups. Overall, the data underscores that the elderly are more susceptible to severe and critical symptoms, making age-specific symptom monitoring essential in predictive healthcare.

4.5. Clustering Analysis (K-Means)

K-Means clustering (k = 5) is utilized to categorize patient profiles into different risk clusters.

Table 3: Cluster Description

Cluster	Profile Description	Disease Likelihood
Cluster 1	Elderly with multiple symptoms and comorbidities	Very High
Cluster 2	Adults with mild symptoms and no comorbidities	Low
Cluster 3	Children with fever and rash	Medium
Cluster 4	Seniors with chronic respiratory issues	High
Cluster 5	Young adults with fatigue and low-grade fever	Medium

The table 3 classifies patient profiles into five different clusters based on symptom patterns, comorbidity status, and demographic factors, each associated with a different likelihood of infectious disease occurrence. Cluster 1, consisting of elderly individuals with multiple symptoms and pre-existing comorbidities, has a very high disease likelihood, denoting extreme vulnerability. Cluster 2 contains adults showing only mild symptoms and no comorbidities, resulting in a low disease likelihood, suggesting a lower probability of severe infection. Cluster 3 represents children presenting with fever and rash, leading to a medium disease likelihood, offering a moderate risk level that warrants monitoring. Cluster 4, seniors with chronic respiratory issues, have a high disease likelihood, underscoring the impact of pre-existing respiratory conditions on infection severity. Cluster 5 explains young adults with fatigue and low-grade fever, also falling into the medium disease likelihood category, denoting moderate but non-negligible risk. Overall, the clustering exposes that elderly patients with multiple health complications face the highest risks, while healthy adults with mild symptoms are at the lowest end of the spectrum.

4.6. Feature Importance from Random Forest

Top 10 predictive features are ranked by importance score represented in Table 4.

Table 4: Feature Importance

Feature	Importance Score
Breathlessness	0.183
Chest Pain	0.161
Fever	0.141
Age	0.132
Comorbidity Count	0.109
Fatigue	0.093
Oxygen Saturation	0.082
Blood Pressure	0.067
Rash	0.054
Cough	0.048

The table 4 ranks features by their importance scores in predicting infectious disease severity within the hybrid machine learning model, revealing which variables most strongly influence classification outcomes. Breathlessness tops the list with an importance score of 0.183, indicating it is the most significant predictor, likely due to its strong association with severe respiratory infections. Chest pain follows closely at 0.161, highlighting its critical role in identifying high-risk cases, especially among older patients. Fever (0.141) and Age (0.132) also rank highly, reflecting their strong correlation with disease occurrence and progression. The Comorbidity count (0.109) underscores the elevated risk posed by pre-existing health conditions. Fatigue (0.093) and Oxygen

saturation (0.082) provide additional diagnostic value, particularly for early detection of deterioration. Blood pressure (0.067) shows a moderate influence, potentially linked to cardiovascular strain during infection. Lower in the ranking are Rash (0.054) and Cough (0.048), suggesting they are less predictive in isolation but still contribute to the overall model's decision-making process when combined with other symptoms. This ranking helps prioritize clinical assessments and supports targeted intervention strategies.

4.7. Statistical Significance Test

To validate the superiority of the proposed Hybrid Machine Learning Model over baseline classifiers such as Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression (LR), a statistical significance test is conducted. The objective is to determine whether the observed performance improvements are genuine or could be attributed to random variation in the dataset.

Hypothesis Formulation: Let M_H denote the performance metric (e.g., Accuracy, Precision, Recall, F1-Score, AUC-ROC) for the proposed Hybrid model, and M_B denote the same metric for a baseline model.

- **Null Hypothesis (H_0) (eq. (11)):**

$$H_0: M_H = M_B \qquad (11)$$

There is no statistically significant difference in performance between the Hybrid model and the baseline model.

- **Alternative Hypothesis (H_1) (eq. (12)):**

$$H_1: M_H > M_B \qquad (12)$$

The Hybrid model outperforms the baseline model with statistical significance.

Significance Level: A significance level of $\alpha = 0.05$ is selected, indicating a 5% tolerance for committing a Type I error (rejecting H_0 when it is true).

Test Selection: Given that the model performance scores are obtained from the same test dataset (paired observations), a paired t-test is employed. This test is appropriate for comparing two related samples where differences in paired values are normally distributed.

The paired t-test statistic is computed using eq. (13).

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad (13)$$

Where:

- $\bar{d}$ = mean of the differences between paired scores
- s_d = standard deviation of the differences
- n = number of paired observations

Decision Rule:

- If p-value < α, reject H_0 and conclude that the Hybrid model significantly outperforms the baseline.
- If p-value ≥ α, fail to reject H_0 and conclude that there is no statistically significant difference.

The paired t-test is conducted for each performance metric comparing the Hybrid model against SVM, Decision Tree, and Logistic Regression.

Table 5: Statistical Test Results

Metric	Hybrid vs SVM (p-value)	Hybrid vs DT (p-value)	Hybrid vs LR (p-value)	Significant at α=0.05?
Accuracy	0.002	0.001	0.004	Yes
Precision	0.015	0.012	0.017	Yes
Recall	0.028	0.019	0.031	Yes
F1 Score	0.009	0.004	0.011	Yes
AUC-ROC	0.006	0.003	0.008	Yes

In Table 5, across all performance metrics, the p-values are below the α threshold of 0.05, indicating that the Hybrid model's performance improvements over the baseline models were statistically significant. This validates that the enhanced predictive capabilities are not due to random variation but are a result of the model's optimized design and feature integration. The difference is statistically significant, validating the improved performance of the hybrid approach.

5. Conclusion and Future Work

In this work, a hybrid model is implemented and evaluated for the precise prediction and risk stratification of infectious diseases in a large-scale healthcare data environment. The incorporation of Random Forest-based feature selection, K-Means clustering for patient profiling, and Gradient Boosting classification are demonstrated to be highly effective in handling high-dimensional, noisy, and heterogeneous medical data. The model is successfully discovered key symptom-disease-age associations, providing interpretable insights for clinicians.

The experimental results are illustrated that the proposed hybrid model is outperformed conventional machine learning algorithms in terms

of accuracy, precision, recall, F1-score, and AUC. Furthermore, the utilization of clustering techniques are permitted the discovery of latent patient groups with similar clinical characteristics, improving personalized risk analysis. Correlation and decision tree-based analyses are discovered age-specific critical symptoms, such as chest pain and breathlessness among elderly patients, supporting proactive and targeted healthcare interventions. Overall, this hybrid model provides a robust, scalable, and explainable solution suitable for real-world infectious disease surveillance systems and hospital decision support tools. Future work will emphasis on integrating temporal and geospatial data, expanding to multi-label disease prediction, and applying deep learning enhancements while maintaining model interpretability.

References

1. N. Sharma, J. Dev, M. Mangla, V. M. Wadhwa, S. N. Mohanty and D. Kakkar, "A Heterogeneous Ensemble Forecasting Model for Disease Predictions", *New Gen Computing*, Vol. 39 (3), pp. 701-715, 2021.
2. S. Grampurohit and C. Sagarnal, "Disease Predictions using Machine Learning Algorithm", *International Conferences for Emerging Technology (INCET)*, Vol. 4 (3), pp. 1-7, 2020.
3. P. Dutta, S. Paul, A. J. Obaid, S. Pal and K. Mukhopadhyay, "Feature Selections based Artificial Intelligence Technique for the Predictions of COVID like Diseases", *2nd International Conference on Physics and Applied Sciences (ICPAS), Journal of Physics: Conference Series*, Vol. 1963 (1), pp. 1-11, 2021.
4. F. E. Ayo, J.B. Awotunde, R.O. Ogundokun, S.O. Folorunso and A.O. Adekunle, "A Decision Support System for Multi-Target Disease Diagnosis: A Bioinformatics Approach", *Heliyon, cell press*, Vol. 6, pp. 1-14, 2020.
5. H. H. Thary and K. Azidan, "A Framework Questionnaire for Diagnosing Infectious Disease using Machine Learning Techniques", INTCSET 2020, *IOP conference series: Materials science and Engineering, IOP Publishing*, Vol. 1094, pp. 1-10, 2021.
6. M. Mariki, E. Mkoba and N. Mduma, "Combining Clinical Symptoms and Patient Features for Malaria Diagnosis: Machine Learning Approach", *Applied Artificial Intelligence, Taylor & Francis,* Vol. 36 (1), pp. 1-25, 2022.
7. Y. A. Adamu and J. Singh, "Malaria Prediction Model using Machine Learning Algorithms", *Turkish journal of computer and mathematics Education*, Vol. 12 (10), pp. 7488-7496, 2021.
8. J. Gao, J. Li and M. Wang, "Time Series Analysis of Cumulative Incidences of Typhoid and Paratyphoid Fevers in China Using both Grey and SARIMA Models", *PLOS ONE,* Vol. 15 (10), pp. 1-14, 2020.
9. H. Wang, N. Wang, M. Li, S. Mi and Y. Shi, "Student Physical Health Information Management Model under Big Data Environment", Hindawi Scientific Programming, Vol. 2021, pp. 1-10, 2021.
10. M. Wang, C. Lee, W> Wang, Y. Yang and C. Yang, "Early Warning of Infectious Diseases in Hospitals based on Multi-Self-Regression Deep Neural Network", Hindawi Journal of Healthcare Engineering, Vol. 2022, pp. 1-13, 2022.
11. M. K. Singh, K. P. Singh and D. Kumar, "Prediction and Detection of Infectious Disease through Machine Learning", European Chemical Bulletin, Vol. 12 (1), pp. 4433-4446, 2023.
12. D. Swain, U. Mehta, A. Bhatt, H. Patel, K. Patel, D. Mehta, B. Acharya, V. C. Gerogiannis, A. Kanavos and S. Manika, "A Robust Chronic Kidney Disease Classifier using Machine Learning", Electronics, MDPI, Vol. 12 (212), pp. 1-13, 2023.
13. Y. Liu, Y. Zhuang, L. Yu, Q. Li, C. Zhao, R. Meng, J. Zhu and X. Guo, "A Machine Learning Framework based on Extreme Gradient Boosting to predict the Occurrence and Development of Infectious Diseases in Laying Hen Farms, Taking H9N2 as an Example", Animals, MDPI, Vol. 13 (1494), pp.1-15, 2023.
14. H. F. Ahmad, H. Khaloofi, Z. Azhar, A. Algosaibi and J. Hussain, "An Improved COVID-19 Forecasting by Infectious Disease Modelling using Machine Learning", Applied Sciences, MDPI, Vol. 11 (11426), pp. 1-38, 2021.
15. M. Hussain, M. A. Cifci, T. Sehar, S. Nabi, O. Cheikhrouhou, H. Maqsood, M. Ibrahim an F. Mohammad, "Machine Learning based efficient Prediction of Positive Cases of

Waterborne Disease", BMC Medical Informatics and Decision Making, Vol. 23 (11), pp. 1-16, 2023.

16. M. Mwamnyange, E. Luhanga and S. R. Thodge, "Big Data Analytics Framework for Childhood Infectious Disease Surveillance and Response System using Modified MapReduce Algorithm", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12 (3), pp. 1-13, 2021.
17. S. Palaniappan, R. V, B. David and P. N. S, "Prediction of Epidemic Disease Dynamics on The Infectious Risk using Machine Learning Algorithms", SN Computer Science, Vol. 3 (47), pp. 1-3, 2022.
18. G. Dhiman, S. Juneja, H. Mahafez, I. E. Bayoumy, L. K. Sharma, M. Hadizadeh, M. A. Islam, W. Viriyasitavat and U. Khandaker, "Federated Learning Approach to Protect Healthcare Data over Big Data Scenario", Sustainability, MDPI,Vol. 14 (2500), pp. 1-14, 2022.
19. M. Uppal, D. Gupta, S. Juneja, A. Sulaiman, K. Rajab, A. Rajeb, M. A. Elmagzoub and A. Shaikh, "Cloud based Fault Prediction for Real Time Monitoring of Sensor Data in Hospital Environment using Machine Learning", Sustainability, MDPI, Vol. 14 (11667), pp. 1-19, 2022.
20. L. B. Amusa, H. Twinomurinzi, E. Phalane and R. N. P. Mafuya, "Big Data and Infectious Disease Epidemiology: Bibliometric Analysis and Research Agenda", Interactive Journal of Medical Research, Vol. 12, pp. 1-16, 2023.
21. S. K. Yadav and Y. Akhter, "Statistical Modelling for the Prediction of Infectious Disease Dissemination with Special Reference to COVID-19 Spread", Frontiers in Public Health, Vol. 9, pp. 1-27, 2021.
22. A. M. Mounir, M. I. Marie and L. A. Elhamid, "Big Data Framework for Predicting Infectious Diseases to Improve Healthcare by Discovering New Symptom Patterns", Journal of Computer Science, Science Publications, Vol. 20 (10), pp. 1-12, 2024.
23. A. P. Zhao, S. Li, Z. Cao, P. J. H. Hu, J. Wang, Y. Xiang, D. Xie and X. Lu, "AI for Science: Predicting Infectious Diseases", Journal of Safety Science and Resilience, Vol. 5, pp. 130-146, 2024.
24. M. M. B. Azam, F. Anwaar, A. M. Khan, M. Anwar, H. B. A. Ghani, T. A. E. Eisa and A. Abdelmaboud, , "A Hybrid Contextual Framework to Predict Severity of Infectious Disease: COVID-19 Case Study", Egyptian Informatics Journal, Vol. 27, pp. 1-13, 2024.
25. V. Nagagopiraju, B. Sujatha, U. Premraja, P. P. Ramaraju and P. Y. Khan, "Prediction of Infectious Disease using AI-Enabled Medical Chatbot Model", International Journal of Communication Networks and Information Security, Vol. 25 (3), pp. 1-6, 2025.
26. S. A. Hudu, A. S. Alshrari, E. J. I. A. Shoura, A. Osman and A. O. Jimoh, "A Critical Review of the Prospect of Integrating Artificial Intelligence in Infectious Disease Diagnosis and Prognosis", Wiley Interdisciplinary Perspectives on Infectious Disease, Vol. 2025, pp. 1-14, 2025.
27. S. A. Ajagbe and M. O. Adigun, "Deep Learning Techniques for Detection and Prediction of Pandemic Disease: A Systematic Literature Review", Multimedia Tools and Applications, Vol. 83, pp. 5893-5927, 2024.
28. F. Ekundayo, "Using Machine Learning to Predict Disease Outbreaks and Enhance Public Health Surveillance", World Journal of Adavanced Research and Reviews, Vol. 24 (03), pp. 794-811, 2024.
29. A. Adeoye, I. A. Onifade, M. Bayode, I. M. Ariyibi, B. Akangbe, O. Akomolafe, T. Ajisafe, D. Hossain and O. F. Owoeye, "Artificial Intelligence and Computational Methods for Modelling and Forecasting Influenza like Illness: A Scoping Review", Beni Suef University Journal of Basic and Applied Sciences, Springer Open, Vol. 14 (93), pp. 1-20, 2025.
30. N. Nuha, S. A. Pitchay, A. H. A. Halim, M. A. B. Sahbudin and L. Sahbudin, "Beyond the Outbreak: A Review of Big Data Analytics in Proactive Infectious Disease Prevention for Risk Mitigation for COVID-19", Journal of Big Data, Springer Open, Vol. 12 (185), pp. 1-23, 2025.
31. D. Xu, W. H. Chan, H. Haron, H. W. Nies and K. Moorthy, "From COVID-19 to Monkeypox: A Novel Predictive Model for Emerging Infectious Diseases", Big Data Mining, Springer Open, Vol. 17 (42), pp. 1-25, 2024.